# Modeling Frequency Data: Methodological Considerations on the Relationship between Dictionaries and Corpora

Karlheinz Mörth, Laurent Romary, Gerhard Budin, Daniel Schopper

**Journal of the Text Encoding Initiative**

**Issue 8 | 2014**
**Selected Papers from the 2013 TEI Conference**

# Modeling Frequency Data: Methodological Considerations on the Relationship between Dictionaries and Corpora

**Karlheinz Mörth, Laurent Romary, Gerhard Budin and Daniel Schopper**

**Publisher**
TEI Consortium

**Electronic version**
URL: http://jtei.revues.org/1356
ISSN: 2162-5603

**Electronic reference**
Karlheinz Mörth, Laurent Romary, Gerhard Budin, and Daniel Schopper, « Modeling Frequency Data: Methodological Considerations on the Relationship between Dictionaries and Corpora », *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015, Online since 21 November 2015, connection on 30 November 2016. URL : http://jtei.revues.org/1356 ; DOI : 10.4000/jtei.1356

# *Modeling Frequency Data: Methodological Considerations on the Relationship between Dictionaries and Corpora*

**Karlheinz Mörth, Laurent Romary, Gerhard Budin, and Daniel Schopper**

## 1. Introduction

1 Academic dictionary writing is making greater and greater use of the TEI Guidelines' dictionary module. And as increasing numbers of TEI dictionaries become available, there is an ever more palpable need to work towards greater interoperability among dictionary writing systems and other language resources that are needed by dictionaries and dictionary tools. In a world of exponentially increasing information, the borders between different types of digital language resources are assuming a role that requires increased attention. Two particularly important instances of such language resources are digital text corpora and dictionaries, both of which play an important part in the TEI community.

2 The research described in this paper has been based on work accomplished in a bundle of linguistically focused projects that—among other activities—also work on glossaries and dictionaries which are intended to be usable both by human readers and by particular NLP applications. The main questions that will be addressed are:

- How can we define the relationship between a dictionary and other language resources such as digital corpora, irrespective of whether they are used in the production of the dictionary or to enrich existing lexicographic data?
- How can this best be documented using the TEI Guidelines?

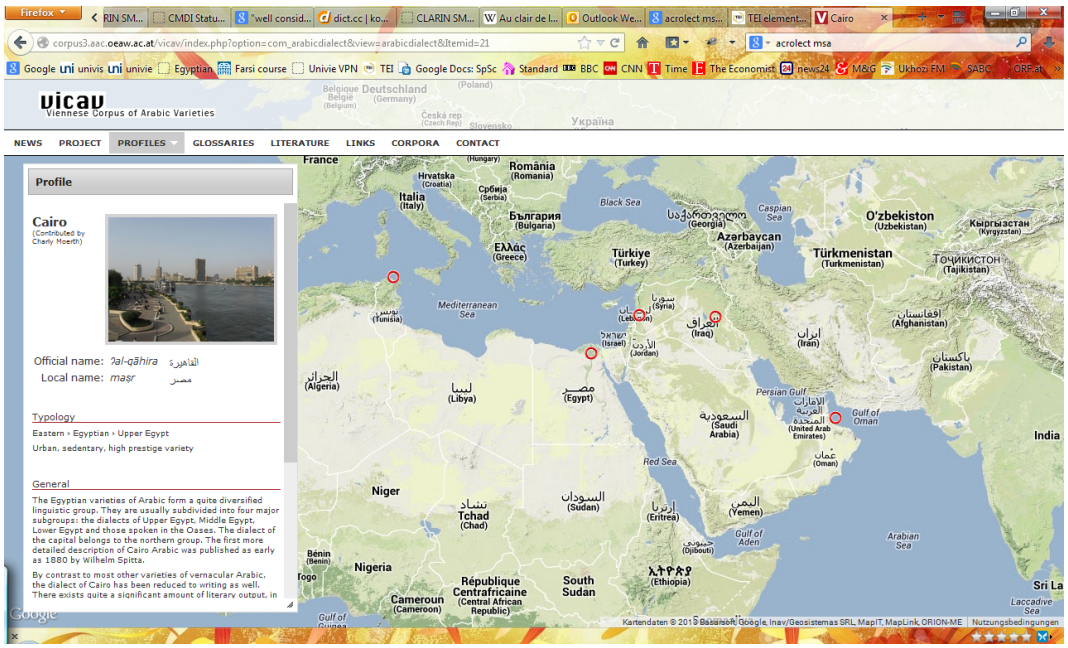
3 The paper comprises two parts: in the first, the authors give a concise overview of the scholarly background of the projects involved and their goals. The second part touches on encoding issues in the related dictionary production. We will focus particularly on the modeling of an encoding scheme for statistical information on lexicographic data gleaned from digital corpora.

## 2. The Dictionaries and Projects Involved

4 The projects in which the dictionaries and related technologies have been developed are tightly interlinked: they are all joint endeavors of the Austrian Academy of Sciences and the University of Vienna, and all conduct research in the field of variational Arabic linguistics. It is important to note that Arabic is characterized by a complex polyglossic situation, with Modern Standard Arabic (MSA) on one side of the spectrum and spoken vernaculars on the other side. Linguists and lexicographers may be confronted with three or even four varieties being used by the same speakers in one and the same linguistic biotope in some regions.

5 The first project to be mentioned is the *Vienna Corpus of Arabic Varieties* (VICAV; see figure 1), which was started two years ago with a low budget, and was intended as an attempt at setting up a comprehensive research environment for scholars pursuing comparative interests in the study of Arabic dialects. The evolving VICAV platform aims at pooling linguistic research data, including various language resources such as language profiles, dictionaries, glossaries, corpora, and bibliographies. One of the main objectives of the project is the creation of a number of dictionaries of Arabic varieties that are primarily intended for comparative purposes.

**Figure 1: A screenshot of a prototype of the VICAV website in 2012.**




6 The second project to be mentioned here is *Linguistic Dynamics in the Greater Tunis Area: A Corpus-based Approach* (TUNICO). This project is financed by a grant from the Austrian Science Fund and aims at the exploration of hitherto poorly-documented contemporary Arabic of the Tunisian capital, which is linguistically and demographically a highly dynamic region. A particular feature of the project is the importance of the dictionary–corpus interface, which will allow the researcher to navigate from the corpus to the dictionary and *vice versa*.The TUNICO project is producing two digital language resources: a corpus of spoken youth language and a diachronic dictionary of Tunisian Arabic. The project started in August 2013 and will run for three years.

7 The third project has grown out of a master's thesis and deals with the lexicographic analysis of the Egyptian vernacular Arabic Wikipedia (Siam 2013). Siam extracted the two hundred most frequent words from *Wikipedia Masri*,[1] which given the scarcity of available tools proved to be quite a challenge. The idea of incorporating statistical information gathered in this project was the initial incentive to start thinking about how to encode such information in accordance with the TEI Guidelines.

8 Four of the dictionaries created in the above-mentioned projects (namely Egyptian, Damascene, Moroccan, and Tunisian) can be correlated with digital corpora that already exist. This does not imply that the existing data have been compiled on the basis of these corpora, none of which are very large. Nonetheless, they are so far the only available digital text collections that can be used to underpin this dictionary-building process with empirical methods. Egyptian is the most widely used Arabic dialect. There is plenty of material on the Internet: of particular interest is the great amount of data that can be found in social media sites and on personal web pages. However, most of this data is of a very hybrid nature and intermixed with MSA. Therefore, it is difficult to use for dialectological research. The only resource we could easily avail ourselves of so far is the Egyptian Wikipedia, which has been made accessible as a corpus as part of another ACDH project working on the conversion of Wikipedias into TEI. This work is particularly interesting for under-resourced languages without other digital texts suitable for linguistic research.

9 VICAV contains a small corpus of samples of Damascene Arabic, which was compiled by Carmen Berlinches during her seven-year stay in Damascus. In addition, there exists the Graz Corpus of Moroccan Arabic, compiled as part of a project funded by the Austrian Science Fund,[2] and the above-mentioned TUNICO corpus which is currently being compiled.[3]

10 Some of these data have already been used to enhance the existing dictionaries, in particular the Egyptian and the Moroccan ones. Many of the words, word forms, and example sentences contained in the corpora have been integrated into the dictionaries. The idea of integrating frequency data grew out of the question as to which lemmas were more important than others.

Table 1: Arabic language corpora

| Source lang. | Target lang. | Corpus | Size (entries) |
|---|---|---|---|
| **Egyptian** | en, de | Wikipedia Masri | ~3,000 |
| **Damascus** | en, de, sp | CB-Corpus | ~3,000 |
| **Tunis** | en, de, (fr) | TUNICO | ~4,000 |
| **Rabat** | en, de | (GCMA) | ~500 |

11 More dictionaries are under preparation. The VICAV database also contains data on Sudanese Arabic, Maltese, Modern Standard Arabic, and the Shawi dialects. One overarching goal of all these endeavors is the creation of a comparative dictionary with an integrated research environment that allows access to all of these data.

## 2.1 A Trimmed Dictionary Schema

12 Using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. Our paper will not reprise the discussion of TEI vs. LMF vs. LEXml vs. Lift vs. RDF vs. other standards;[4] we assume that the TEI dictionary module is sufficiently well developed to cope with all requirements of our projects. The basic schema used has already been tested in several projects for various languages and will furnish the foundation for the intended customizations.

13 Created to serve as sources for comparative research, all of the above-mentioned dictionaries have to fulfill a series of requirements: Technically, they have to be processable by various tools, most importantly by several web services on which the dictionary tools build. They have to be compatible with one another and the tools used in their creation. Therefore, they have to be encoded following one single schema in order to allow electronic tools to work on them in tandem and to allow users to execute meaningful queries across all of the dictionaries. This goal has so far been achieved by applying a narrowly defined schema that imposes a number of specific constraints, which were meant to serve as a mechanism to enhance interoperability. In all design issues we have strived for a high degree of compliance with LMF. The main methods of imposing such constraints are reducing alternate constructs and matching with constructs of LMF (ISO 2008).[5]

14 Since all of these data are "born digital," it is comparatively easy to ensure the structural uniformity of the dictionaries. Basically, our dictionaries are conceptualized as a specific type of text and are therefore encoded with `<text>` elements. Each dictionary starts with a `<teiHeader>` which contains the metadata of the dictionary. The `<body>` of the VICAV dictionaries contains two `<div>` elements: one typed `"entries"`, holding all entries of the dictionary, and another typed `"examples"`, which is populated with a series of `<cit>` / `<quote>` constructs containing example

sentences.[6] Treating these examples as independent units allows dictionary writers to reuse the same sentence in various parts of a dictionary. The schema uses the <entry> element, does not make use of the <hom> element, and does not allow <superEntry>, <entryFree>, or <dictScrap>.[7]

15 Thus, our TEI dictionaries basically look like this:

```
<TEI>
 <body>
  <div type="entries">
   <entry>...</entry>
   <entry>...</entry>
   <entry>...</entry>
  ...
  </div>
  <div type="examples">
   <cit type="example">...</cit>
   <cit type="example">...</cit>
   <cit type="example">...</cit>
  ...
  ...
  </div>
 </body>
</TEI>
```

16 A typical, slightly simplified <entry> taken from the Egyptian dictionary is shown below:

```
<entry xml:id="kitaab_001">
 <form type="lemma">
  <orth xml:lang="ar-arz-x-cairo-vicav">kitāb</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">كتاب</orth>
 </form>

 <gramGrp>
  <gram type="pos">noun</gram>
  <gram type="root" xml:lang="ar-arz-x-cairo-vicav">ktb</gram>
 </gramGrp>

 <form type="inflected" ana="#n_pl">
  <orth xml:lang="ar-arz-x-cairo-vicav">kutub</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">كتب</orth>
 </form>

 <sense>
  <cit type="translation" xml:lang="en">
   <quote>book</quote>
  </cit>

  <cit type="translation" xml:lang="de">
   <quote>Buch</quote>
  </cit>
 </sense>
</entry>
```

17 To minimize processing overhead, hierarchical nesting is avoided whenever possible; even inflected forms of the lemma are treated like other variant word forms and are encoded as `<form>` elements on the same hierarchical level as the lemma itself.

18 The above structure is the one actually implemented at the moment. All VICAV dictionaries and most other dictionaries being produced at the ACDH follow this basic structure.[8]

## 2.2 Frequency Data

19 Lexicostatistical data and methods are used in many fields of modern linguistics, lexicography being only one of them. Modern dictionary production relies on corpora, and statistics play an important role in lexicographers' decisions, for instance when selecting lemmas to be included in dictionaries or selecting senses to be incorporated into dictionary entries. However, lexicostatistical data are not only of interest for the lexicographer; they might also be useful to the users of lexicographic resources, especially digital lexicographic resources. The question as to how to make such information available takes us to the issue of how to encode it.

20 Reflecting on the dictionary–corpus interface and on the issue of how to bind corpus-based statistical data into the lexicographic editing workflow, two prototypical approaches are conceivable: (1) either statistical information is statically embedded in the dictionary entries or (2) a dictionary interface provides functionalities to access services capable of providing the required data.

21 A group of people working on methodologies to implement functionalities of the second type is the Federated Content Search (FCS) working group, an initiative of the CLARIN-ERIC infrastructure which strives to enhance search capabilities in locally distributed data stores (Stehouwer et al. 2012). FCS is intended to work with heterogeneous data, and dictionaries are only one type of language resource to be taken into consideration. In view of the growing prevalence of more dynamic digital environments, the second of the above-mentioned approaches is more appealing. In practice, the digital workbench will require both options. This is particularly true given that corpora change and grow over time. Resolving polysemy and grouping instances into senses remain tasks that cannot be achieved automatically—yet for the sake of verifiability these should be as accountable as possible. The prevalence of digital workflows in lexicographic editing in combination with the availability of large-scale data storage at reasonable cost provide the technological prerequisites to envision systems that keep track of such editorial processes and make the lexicographers' decisions more transparent and verifiable.

## 3. Documenting Consulted Corpora

22 If traceability is to be one of the fundamental benefits of a digital lexicographic workflow, documenting the provenance of the language data which editorial decisions rely upon becomes a basic requirement. Among the various possible elements to accommodate such metadata in the current TEI Schema, the dictionary's <sourceDesc> element with <bibl>elements (or their finer-grained variants <biblStruct> or <biblFull>) might seem an obvious fit.

```
<sourceDesc>
 <bibl>Österreichisches Wörterbuch. 42. Auflage. Wien: ÖBV 2012</bibl>
 <bibl>amc – Austrian media Corpus. Austrian Centre for Digital Humanities.
Austrian Academy of Sciences http://www.oeaw.ac.at/acdh/amc</bibl>
</sourceDesc>
```

23 This solution, however, is far from ideal, at least in cases where the digital dictionary stems from a printed original, as their different relations with respect to the <text> would be indiscernible in the markup. Adding @type attributes to both <bibl> elements would not help either, as that would merely classify the bibliographic records, not the function of the entities they describe.[9] Thus, the distinction between two kinds of "sources" should be made on a more general level. The following construct would be formally valid:

```
<sourceDesc>
 <listBibl type="printedSource">
  <bibl>Österreichisches Wörterbuch. 42. Auflage. Wien: ÖBV 2012</bibl>
 </listBibl>

 <listBibl type="consultedCorpora">
  <bibl>amc – Austrian media Corpus. Austrian Centre for Digital Humanities.
Austrian Academy of Sciences http://www.oeaw.ac.at/acdh/amc</bibl>
 </listBibl>
</sourceDesc>
```

24 Grouping those two kinds of bibliographic references in separate listBibl elements with appropriate @type attributes—for instance "printedSource" and "consultedCorpora"—solves the issue of ambiguity at least on the surface. There remain concerns, however. First of all,

relying on attribute values on parallel constructs to encode such a fundamental difference is not as expressive as one might wish—especially in the case of a much-used attribute like `@type`. More importantly, does this kind of markup (or any other of the approaches mentioned above) sufficiently denote the specific type of relationship between the digital dictionary and a corpus, which obviously cannot be reduced to one of a "source" and its "derivation?" This conceptual problem pertains to nearly all possible solutions we could think of.

25 One solution would be embedding this kind of metadata into the `<editorialDecl>` element, which "provides details of editorial principles and practices applied during the encoding of a text,"[10] or even into `<samplingDecl>`, where "the rationale and methods used in selecting texts, or parts of text, for inclusion in the resource"[11] are documented. Although this seems more accurate with respect to the role of language resources in dictionary writing, the definition of their common ancestor `<encodingDesc>` ("documents the relationship between an electronic text and the source or sources from which it was derived")[12] explicitly refers to a relation of dependency of one on the other, which seems inappropriate in our case.

26 Surprisingly, it is the Critical Apparatus module which provides an appropriate solution for our problem. Originating in the tradition of textual criticism, this module defines the phrase-level element `<app>` to embed various versions of a passage in-line, optionally declaring one as the preferred reading. The resulting TEI `<text>` is a compound object that does not have one single "non-electronic" counterpart, but documents a multitude of fragments from various resources. Likewise, a dictionary, which is closely intertwined with data from language resources, can be conceptualized as the abstraction of this instance data. In order to express the specific nature of its "sources," the `textcrit` module has to define a new child to `<sourceDesc>`, the `<listWit>` element.

27 Following this example, we propose the introduction of a `<listResource>` element to hold a list of any language resources from which the dictionary in the document's `<body>` draws its statistical information. This list consists of one or more `<resource>` elements, which provide relevant metadata about each language resource and include pointers to its content.

- `<resource>` describes a language resource of any kind (including, but not limited to, text corpora) that has been used as source material in the creation of a dictionary.
- `<listResource>` (language resource list) contains a list of language resources of any kind that have been used as source material in the creation of a dictionary.

28 By choosing a deliberately broad term as the new element's name, we try to keep the range of possible language resource types as open as possible without confining their possible function to that of a statistical source.

```
<teiHeader>
 <fileDesc>
  <sourceDesc>
   <bibl>Österreichisches Wörterbuch ...</bibl>
   <acdh:listResource>
    <acdh:resource xml:id="corpus1">...</acdh:resource>
    <acdh:resource xml:id="corpus2">...</acdh:resource>
    <acdh:resource xml:id="corpus3">...</acdh:resource>
   </acdh:listResource>
  </sourceDesc>
 ...
 </fileDesc>
</teiHeader>
```

29 Making `<listResource>`a child element of `<sourceDesc>` tries to address both kinds of corpus–dictionary relations we have come across in our projects: it takes into account cases where a "born-digital" dictionary draws most of its material from language resources—making them thus an important, but still intermediate source—while possibly drawing a clear line between a source of a dictionary's `<text>` and the source of statistical data the `<text>` has been enriched with.

30 The purpose of the `<resource>` element is twofold: first of all, it enables a user to locate and access the language resource personally. Secondly, it provides a basic description of it through a series of appropriate properties. Although this information is likely to be part of the metadata held alongside the corpus data itself, it seems reasonable to keep a summary of it with the dictionary, especially as corpora may become inaccessible or evolve over time. The TEI Guidelines already provide the components for this in the `<bibl>` element's content model, with the `<extent>`/`<measure>` construct being a natural candidate for the representation of corpus size.

```
  <acdh:resource xml:id="amc" type="digitalcorpus" subtype="media"
status="closed">
   <title>Austrian Media Corpus</title>
   <ptr target="http://acdh.oeaw.ac.at/corpora/amc"/>
   <publisher>
    <name>Austrian Centre for Digital Humanities. Austrian Academy of
Sciences</name>
    <address>
     <street>Sonnenfelsgasse 19/8</street>
     <postCode>1010</postCode>
     <name type="settlement">Vienna</name>
     <name type="country">Austria</name>
    </address>
   </publisher>
   <date type="content" notBefore="1987" notAfter="2012">1987–2012</date>
   <date type="published">2013</date>
   <date type="lastAccess">2014-01-30</date>
   <extent>
    <measure commodity="tokens" quantity="8512255860"/>
    <measure commodity="words" quantity="6228727272"/>
    <measure commodity="sentences" quantity="425830965"/>
    <measure commodity="paragraphs" quantity="61235319"/>
    <measure commodity="documents" quantity="33662024"/>
   </extent>
   <note type="description">Contains articles from Austrian newspapers and
magazines between 1987 and May 2012. PoS tagging with “Tree Tagger” and
“RFTagger”</note>
  </acdh:resource>
```

31 Modeling `<resource>` after `<bibl>` helps us address some specific limitations of the language resources we currently have to make do with: it lets us distinguish between various kinds of data via the `@type` and `@subtype` attributes, while the `@status`attribute indicates whether the data in question are expected to be subject to change or not. Since we consider it important to document the essential properties of a corpus (status, size, date of its content) in a consistent manner, we decided to narrow down the possible components of `<bibl>` significantly and create a specially tailored version from it.

32 Many other aspects of language resources may be desirable to record—especially considering questions of mid- and long-term preservation. Since data used in lexicographic production are most likely to be distributed over a wide range of organizations and locations, accessibility cannot be assumed in all cases. In order for a user to be able to assess the relevance of any language resource in the context of the final dictionary, a multitude of parameters has to be taken into consideration. For instance, it would be important to identify inherent biases in the sociologic, geographic, or diachronic sampling of language data or technical limitations in its markup. The German LAUDATIO Project has developed a comprehensive, `<teiHeader>`-based corpus description specification which is directly aimed at this purpose.[13] In light of such issues it seems advisable to allow the inclusionof the `<teiHeader>` of a corpus (or possibly any descriptive metadata) inside our `<resource>` element, to document its state at the time of the dictionary's publication.

# 4. Documenting Corpus Queries

33 So far, we have defined a way to describe the language resources we want to refer to in the dictionary's `<teiHeader>`. This leaves the following questions to be addressed: which information is to be encoded when documenting our corpus queries? Which parts of those entries do we need to attach this information to? And, finally, how can we establish the linkage between our description of the corpus instance data, the dictionary, and the corpus metadata held in the `<listResource>`element?

## 4.1 The Tenets of Frequency Information

34 What is needed is a definitive system to register quantifications of particular items represented in dictionary entries. This of course raises the question as to which parts of a dictionary entry can be considered relevant. First to come to mind, of course, are headwords. But there are many other constituents of dictionary entries that might be furnished with frequency data: inflected word forms, collocations, multiword units, and particular senses are relevant items in this respect.

35 The data model should not only provide elements to encode frequencies within elements describing the above-listed constituents of entries, but also allow indication of the source from which the data were gleaned and how the statistical information was created. The basic constituents of our model should contain these items:

- Value (number of occurrences of the particular item)
- Rank
- Provenance (source from which the data is taken)
- Retrieval method (how the statistical information was created)
- Query type
- Evaluation mode

36 Ideally, persistent identifiers should be used to identify not only the corpora but also the services involved in creating the statistical data.

## 4.2 Spoilt for Choice

37 In our attempts to design a viable solution to our encoding problems, we went through three stages: (1) we tried to make use of some TEI elements with very flexible semantics and to provide them with `@type` attributes; (2) we tried to apply TEI feature structures; and (3) we started to work on a new customization.

### 4.2.1 Catch-all Elements

38 As is well known, there are some TEI elements which can be used for almost anything by furnishing them with `@type` attributes. The most commonly used ones are `<note>`, `<ab>`, and `<seg>`, which readily lend themselves to purposes such as ours through their very general semantics. Early attempts of ours to model frequency data also made use of `<list>` and `<item>` elements, resulting in constructs such as those in the example below:

```
<entry xml:id="mashcal_001">
 <form type="lemma">
  <orth xml:lang="ar-arz-x-cairo-vicav">maš ʕal</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">مشعل</orth>

  <list type="statData">
   <item>
    <ref target="http://acdh.oeaw.ac.at/wikiMasri"/>
    <measure type="lemmaNumber">6</measure>
    <measure type="rank">2456</measure>
   </item>
  </list>
 </form>

 <gramGrp>
  <gram type="pos">noun</gram>
  <gram type="root" xml:lang="ar-arz-x-cairo-vicav">š ʕl</gram>
 </gramGrp>

 <form type="inflected" ana="#n_pl">
  <orth xml:lang="ar-arz-x-cairo-vicav">mašāʕil</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">مشا عل</orth>

  <list type="statData">
   <item>
    <ref target="http://acdh.oeaw.ac.at/wikiMasri"/>
    <measure type="tokenNumber">2</measure>
    <measure type="rank">23765</measure>
   </item>
  </list>
 </form>
</entry>
```

39 In this example, the statistical information indicated by means of `<item>` elements refers to the `<form>` elements. This attempt soon seemed unsatisfactory: although sufficiently versatile in its content, `<list>` is intended only to be placed in specific “wrapper” elements of the dictionary module (`<entry>`, `<form>`, and`<sense>`amongst others) but is disallowed in a lot of other contexts

where frequency information is potentially relevant. In particular, this would have excluded the possibility to embed frequency data in elements containing grammatical information (<gram> and its syntactic-sugar equivalents <case>, <gen>, <number>, etc.) as well as <usg> and <def>. The other constructs mentioned above proved equally problematic: the definition of <seg> ("represents any segmentation of text below the 'chunk' level ")[14] hardly permits arbitrary data structures like the ones we needed; <note>, on the other hand, was too likely to semantically interfere with editorial notes (footnotes, marginal notes) in a retro-digitized dictionary; and <ab> would have forced us to use bulky pointing mechanisms to express the relationship between the various parts of an entry and the attached frequency information, since it is only allowed as a sibling of <entry>.

### 4.2.2 Feature Structures

40 In a second approach, we used feature structures, a very versatile, sufficiently well-explored tool for formalizing all kinds of linguistic phenomena. One of the advantages of the <fs> element is that it can be placed inside most elements used to encode dictionaries.

```
<entry xml:id="mashcal_001">
 <form type="lemma">
  <orth xml:lang="ar-arz-x-cairo-vicav">maššal</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">مشعل</orth>

  <fs type="corpFreq">
   <f name="corpus" fVal="#wikiMasri"/>
   <f name="frequency"><numeric value="6"/></f>
   <f name="rank"><numeric value="2456"/></f>
  </fs>
 </form>

 <gramGrp>
  <gram type="pos">noun</gram>
  <gram type="root" xml:lang="ar-arz-x-cairo-vicav">šʕl</gram>
 </gramGrp>

 <form type="inflected" ana="#n_pl">
  <orth xml:lang="ar-arz-x-cairo-vicav">mašāʕil</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">مشا عل</orth>

  <fs type="corpFreq">
   <f name="corpus" fVal="#wikiMasri"/>
   <f name="frequency"><numeric value="2"/></f>
   <f name="rank"><numeric value="23765"/></f>
  </fs>
 </form>
</entry>
```

41 Feature structures are impressively simple to use and have a lot of advantages when it comes to modeling abstract structures and their relations. However, they are not very human-readable, nor are frequencies a "feature" of the entry's components in the strict sense of the word.

### 4.2.3 Attempting Customization

42 All these "conservative" attempts adopted existing elements and resulted in solutions which appeared to be far from perfect, especially `<item>` and `<fs>` being void of relevant semantics. This —in the end—made us think of alternatives by customizing our dictionary scheme and adding a

set of objects (attributes and elements) to describe frequencies in context. With a wide range of different application scenarios in mind, we attempted to design something like a statistical crystal that could also be reused beyond our particular projects.

43 We named the root element to carry frequencies "statistical information." We chose a generic name rather than making use of narrower terms such as "corpusFrequency," as the data we wanted to use might come from other language resources than text corpora, such as word lists, other dictionaries, or databases aggregating statistical data from external sources. The chosen term appeared to be semantically correct and would allow us to keep options open to other scenarios.

- `<statInfo>` (statistical information) contains statistical information about instances of any component of a dictionary entry in one or more language resources.

44 The next step was to find a way to indicate where the statistical information came from. Intuitively, one might expect a `<source>` element. However, `<source>` already exists in the Manuscript Description module (and can only be used as a child of a `<recordHist>` element). As we considered it good practice to avoid denominational ambiguities, we eschewed using "source" in the namespace of the customization, but introduced a `<dataset>` element which, through membership of the `att.canonical` class, inherits the attributes `@key` and `@ref`, with the latter providing the mechanism to point to a `<resource>` element in the `<teiHeader>`.

```
<entry xml:id="mashcal_001">
 <form type="lemma">
  <orth xml:lang="ar-arz-x-cairo-vicav">maš ʕal</orth>
  <orth xml:lang="ar-arz-x-cairo-arabic">مشعل</orth>

  <acdh:statInfo>
   <acdh:dataset ref="#wikiMasri"> ... </acdh:dataset>
  </acdh:statInfo>
 ...
 </form>
</entry>
```

45 In the field of our research, we are still far away from reliable reference corpora in the proper sense of the word. At the moment, we are instead in a situation where we have to integrate anything available in default of anything better. For comparative purposes it might therefore be important to have a list with several `<dataset>` elements to give users a more complete picture of the available data beyond the resource proper.

46 The statistical information itself would remain in the TEI namespace. This is exactly the same construct which we have already proposed above.

```
<measure commodity="tokens" quantity="6" unit="count" type="absolute"/>
<measure type="rank" quantity="2456"/>
```

47 A key issue here is the access mode. The element `<retrievalMethod>` has been proposed to accommodate information regarding the query and possible modifications to the result set. This information is dealt with in two child elements: `<query>` and `<evalMode>`.

```
<acdh:retrievalMethod>
 <acdh:query type="CQP">lemma="go"</acdh:query>
 <acdh:evalMode>manual</acdh:evalMode>
</acdh:retrievalMethod>
```

48 The `<query>` element contains the query string. It should also have a `@type` attribute indicating the applied query language. In the example above, CQP (Corpus Query Processor) refers to the query language of the IMS[15] Corpus Workbench. The element `<evalMode>` can be filled with either `"none"`, which implies that the data was retrieved automatically, or `"manual"`, which should be applied when some kind of postprocessing has been done.

49 For purposes of reproducibility it may be desirable to document the various steps that produced the final set of records with finer granularity. In this case, `<evalMode>` could be replaced by an `<evalDesc>` element, containing a series of `<filter>` tags with one child `<query>` each.

```
  <acdh:retrievalMethod>
   <acdh:query type="CQL">word=".*lein"</acdh:query>
   <acdh:evalDesc>
    <acdh:filter n="1" type="negative" selectedToken="first" from="0" to="0"
includeKwic="true">
     <note>Filtering out named entities</note>
     <acdh:query type="CQL">[pos="NE"]</acdh:query>
    </acdh:filter>
    <acdh:filter n="2" type="negative" selectedToken="first" from="0" to="0"
includeKwic="true">
     <note>Filtering out adjectives and adverbs</note>
     <acdh:query type="CQL">[pos="ADJ|ADJD|ADV"]</acdh:query>
    </acdh:filter>
   </acdh:evalDesc>
  </acdh:retrievalMethod>
```

50 The construct above indicates that the results of the original query (`'word=".*lein"'`) were first reduced by excluding any named entities (`<filter>` number 1), and the resulting subset narrowed down by removing any adjectives and adverbs (`<filter>` number 2). The frequency data in a subsequent `<measure>` element would relate to the final set of records in `<evalDesc>`.

51 Of course, this kind of detail is not attainable without the help of specialized software. Aiming toward tighter integration of language resources with dictionaries, we imagine a next generation of dictionary editors providing facilities to query language resources and keep track of user-driven modifications of the results, and offering functions to embed them in the markup. Until implementations have reached this level of integration, we have to rely on a combination of components to support this functionality. An example for this is the commercial product Sketch Engine,[16]which includes a web interface for querying language corpora. In particular, it offers the ability to download the resulting frequency data (as well as concordances) in its own, vendor-specific, XML format. This format also contains the sequence of query expressions that leads to the final result set. Thus, a simple XSL stylesheet can be used to transform this into the format we have proposed above.

```
<frequency>
 <heading>
  <corpus>amc</corpus>
  <query>word,[word=".*lein"] 0 0 1 [pos="NE"]0 0 1 [pos="ADJ|ADJD|ADV"]</query>
 </heading>
 <block>
  <name>word</name>
  <items>
   ...
   <item>
    <str>fräulein</str>
    <freq>23515</freq>
   </item>
  ...
  </items>
 </block>
</frequency>
```

52 A sufficiently descriptive markup for query results, however, is quite verbose by nature. Depending on the context this practice can have undesirable consequences; for instance, the XML data may become hard to read or, more seriously, whitespace issues may be introduced. If, for example, a dictionary writer states that an entry's headword can occur in two orthographic forms and wants to underpin this assertion with frequency data from a corpus, he or she would have to use a construct like the following, leaving it technically unclear where the headword ends and the statistical metadata begins.

```
<form type="lemma">
 <orth type="main">eintepschen
  <acdh:statInfo>
   <acdh:dataset ref="#amc">
    <acdh:retrievalMethod>
     <acdh:query>[word=".*tepsch.*"]</acdh:query>
     <acdh:evalDesc>
      <acdh:filter type="negative" from="0" to="0" includeKwic="true">
       <acdh:query>[word="Step.*"]</acdh:query>
       <acdh:note>excluding composites with “Step”</acdh:note>
      </acdh:filter>
     </acdh:evalDesc>
    </acdh:retrievalMethod>
    <measure quantity="35" commodity="tokens" type="absolute"/>
    <measure quantity="0.0" commodity="tokens" type="relative"/>
   </acdh:dataset>
  </acdh:statInfo>
 </orth>
 <orth type="variant">eindepschen
  <acdh:statInfo>
   <acdh:dataset ref="#amc">
    <acdh:retrievalMethod>
     <acdh:query>[word=".*depsch.*"]</acdh:query>
     <acdh:evalMode>none</acdh:evalMode>
    </acdh:retrievalMethod>
    <measure quantity="6" commodity="tokens" type="absolute"/>
    <measure quantity="0.0" commodity="tokens" type="relative"/>
   </acdh:dataset>
  </acdh:statInfo>
 </orth>
</form>
```

53 In order to avoid ambiguities, we propose an alternative encoding style: instead of specifying the relation between the instance data and the lexicographic description by nesting the former inside the latter, we suggest using the `linking` module's `@corresp` attribute to point from the

dictionary to a `<statInfo>` element which can be kept in the dictionary's back matter or in another document instance. Although this might add some processing overhead to an application, it improves maintainability and readability significantly by dividing the two layers of information.

```
<entry>
 <form type="lemma">
  <orth type="main" corresp="#si1">eintepschen</orth>
  <orth type="variant" corresp="#si2">eindepschen</orth>
 </form>
...
</entry>
...
<back>
 <acdh:statInfo xml:id="si1">
  <acdh:dataset ref="#amc">
   <acdh:retrievalMethod>
    <acdh:query>[word=".*tepsch.*"]</acdh:query>
    <acdh:evalDesc>
     <acdh:filter type="negative" from="0" to="0" includeKwic="true">
      <acdh:query>[word="Step.*"]</acdh:query>
      <acdh:note>excluding composites with “Step”</acdh:note>
     </acdh:filter>
    </acdh:evalDesc>
   </acdh:retrievalMethod>
   <measure quantity="35" commodity="tokens" type="absolute"/>
   <measure quantity="0.0" commodity="tokens" type="relative"/>
  </acdh:dataset>
 </acdh:statInfo>
 <acdh:statInfo xml:id="si2">
  <acdh:dataset ref="#amc">
   <acdh:retrievalMethod>
    <acdh:query>[word=".*depsch.*"]</acdh:query>
    <acdh:evalMode>none</acdh:evalMode>
   </acdh:retrievalMethod>
   <measure quantity="6" commodity="tokens" type="absolute"/>
   <measure quantity="0.0" commodity="tokens" type="relative"/>
  </acdh:dataset>
 </acdh:statInfo>
</back>
```

54 To take all of this further, we will first create more real-world data with the customized schema and test them in applications currently under development. In addition, it will be necessary to keep discussing the customized dictionary schema with the community. A final goal (the realization of which is admittedly not yet near at hand) is the integration of a viable solution into the TEI Guidelines.

## 5. Conclusions

55 A major interest that has accompanied our experiments is the clearly discernible phenomenon of blurring boundaries between digital language resources. Data available in one resource can be integrated into others; creating new resources from pre-existing ones has become much more feasible. We strongly believe that the permeability between language resources will also change our way of how we look at corpora and dictionaries. In the digital world the two grow closer and closer. Not only do they depend on each other (dictionaries need corpora to be compiled, while corpus tools need dictionaries for annotation); users will increasingly want to use them together, ideally in the same interfaces.

56 Some of the problems described in this paper have not been dealt with so far because readily and freely accessible language resources are not as abundant as one might assume. However, funding agencies increasingly insist on open access not only to research results but also to research data. It is therefore to be hoped that the situation with respect to openly accessible lexicographic data as well as to electronic corpora will improve in the years to come. Solutions for integrating these data and/or accessing various resources simultaneously will become even more important. Thinking about how particular language resources can interact and working on appropriate interfaces is an indispensable prerequisite for more linked (open) data and service-based architectures. The more such data become available, the more important it will become for the TEI to provide viable solutions for dealing with them.

## BIBLIOGRAPHY

Budin, Gerhard, Stefan Majewski, and Karlheinz Mörth. 2012. "Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries." *Journal of the Text Encoding Initiative* 3. http://jtei.revues.org/522. doi:10.4000/jtei.522.

ISO (International Organization for Standardization). 2008. *Language Resource Management — Lexical Markup Framework (LMF)*. ISO 24613:2008. Geneva: ISO.

Romary, Laurent. 2013. "TEI and LMF Crosswalks." In *Digital Humanities: Wissenschaft vom Verstehen* (forthcoming), edited by Stefan Gradmann and Felix Sasaki. Humboldt Universität zu Berlin. http://hal.inria.fr/hal-00762664.

Siam, Omar. 2013. "Ein digitales Wörterbuch der 200 häufigsten Wörter der Wikipedia in ägyptischer Umgangssprache." MPhil thesis, University of Vienna. http://othes.univie.ac.at/26036.

Stehouwer, Herman, Matej Durco, Eric Auer, and Daan Broeder. 2012. "Federated Search: Towards a Common Search Infrastructure." In *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation (LREC 2012)*, edited by Nicoletta Calzolari et al., 3255–59. European Language Resources Association. http://hdl.handle.net/11858/00-001M-0000-000F-9E1D-8.

TEI Consortium. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.6.0. Last updated January 20. http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/.

## NOTES

**1** Masri Wikipedia Introduction in English

**2** *Arabic in the Middle Atlas Mountains (Morocco)*, University of Graz, FWF Project Number P 21722-G20. See http://sprachausbau.uni-graz.at/de/forschen/sprachen-in-nordafrika-marokko/beschreibung/.

**3** The first campaign was undertaken in 2013. The two investigators recorded some 22 hours of audio material, which has been transcribed and are being analyzed.

**4** For a concise overview of these formats compare section 3 "Data Formats" in Budin, Majewski, and Mörth 2012 (http://jtei.revues.org/522#tocto1n3).

**5** See Budin, Majewski, and Mörth 2012 and Romary 2013b.

**6** While entries and example sentences were originally located directly inside the `<body>`, we have now settled for a clearer distinction between the two. Since most of our dictionary data is created and maintained using the Viennese Lexicographic Editor and held in a relational database (described in Budin, Majewski, and Mörth 2012), format changes like this can be applied transparently on all of our dictionaries, making it easy to ensure structural homogeneity.

**7** For more detail see Budin, Majewski, and Mörth 2012.

**8** The example above exhibits some features which might seem idiosyncratic at first glance, namely the usage of the `@ana` attribute and the values in the `@xml:lang` attributes. The fragment identifier in `@ana` refers to a feature library in the `<teiHeader>`, providing a concise notation for morphosyntactic annotations. In the example above, it defines the inflected `<form>` to bear the features *noun* and *plural*. The composition of the `@xml:lang` attributes is an extension to the BCP 47 standard tags, which proved necessary in order to provide a higher degree of locational granularity (see Budin, Majewski, and Mörth 2012).

**9** One might wish to have a `@role` attribute at hand to express this differentiation, yet given its already fairly vague semantics that does not seem advisable either. While `@role` is predominantly used in the realm of names and named entities, it is also defined in `att.tableDecoration`, where it indicates whether a cell holds actual data or just a label. Defining it locally on `<bibl>` would only overload it with another, highly specialized sense and seems a makeshift strategy.

**10** TEI Consortium 2014: Appendix C Elements, http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/ref-editorialDecl.html.

**11** TEI Consortium 2014: Appendix C Elements, http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/ref-samplingDecl.html.

**12** TEI Consortium 2014: Appendix C Elements, http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/ref-encodingDesc.html.

**13** http://www.laudatio-repository.org/repository/files/documentation/corpus/teiODD_LAUDATIODocumentation_S6.zip.

**14** TEI Consortium 2014: Appendix C Elements,http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/de/html/ref-seg.html.

**15** Institut für Maschinelle Sprachverarbeitung, Stuttgart.

**16** https://www.sketchengine.co.uk.

## ABSTRACT

Academic dictionary writing is making greater and greater use of the TEI Guidelines' dictionary module. And as increasing numbers of TEI dictionaries become available, there is an ever more palpable need to work towards greater interoperability among dictionary writing systems and other language resources that are needed by dictionaries and dictionary tools. In particular this holds true for the crucial role that statistical data obtained from language resources play in lexicographic workflow—a role that also has to be reflected in the model of the data produced in these workflows. Presenting a range of current projects, the authors address two main questions in this area: How can the relationship between a dictionary and other language resources be conceptualized, irrespective of whether they are used in the production of the dictionary or to enrich existing lexicographic data? And how can this be documented using the TEI Guidelines? Discussing a variety of options, this paper proposes a customization of the TEI dictionary module that tries to respond to the emerging requirements in an environment of increasingly intertwined language resources.

## INDEX

**Keywords:** lexicography, language resources, digital corpora, statistics

## AUTHORS

### KARLHEINZ MÖRTH

Karlheinz Mörth is currently director of the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences, senior researcher and project leader, lecturer at the University of Vienna, head of the DARIAH National Coordinator Committee and vice-chair of the CLARIN Standards Committee. With a broad background in cultural, literary, and linguistic studies, he has worked on a number of scholarly digital projects. He has contributed to the design and creation of a number of digital language resources, taking responsibility for text encoding and software development. His current research activities focus on eLexicography, text technology for linguistic research and standards for digital language resources.

### LAURENT ROMARY

Laurent Romary is Directeur de Recherche for INRIA (France) and a guest scientist at Humboldt University (Berlin, Germany). He carries out research on the modeling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He received a PhD in computational linguistics in 1989 and his

Habilitation in 1999. He launched and directed the Langue et Dialogue team at Loria (Nancy, France) and participated in several national and international projects related to the representation and dissemination of language resources and to human–machine interaction, coordinating the MLIS/DHYDRO, IST/MIAMM, and eContent/Lirics projects. He has been the editor of ISO standard 16642 (TMF – Terminological Markup Framework) and is the chairman of ISO committee TC 37/SC 4 on Language Resource Management, as well as member (2001–2007), then chair (2008–2011), of the TEI Council. In recent years, he has led the Scientific Information directorate at CNRS (2005–2006) and established the Max-Planck Digital Library (Sept. 2006–Dec. 2008). He is currently director of DARIAH-EU.

### GERHARD BUDIN

Gerhard Budin is a full professor of terminology studies and translation technologies at the Centre of Translation Studies at the University of Vienna, former director of the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences, member (kM) of the Austrian Academy of Sciences, and holder of the UNESCO Chair for Multilingual, Transcultural Communication in the Digital Age. He also serves as vice president of the International Institute for Terminology Research and chair of a technical subcommittee in the International Standards Organization (ISO) focusing on terminology and language resources (ISO/TC 37/SC 2 2001–2009, SC 1 2009–present). His main research interests are language technologies, corpus linguistics, and knowledge engineering, E-Learning technologies and collaborative work systems, distributed digital research environments, terminology studies, ontology engineering, cognitive systems, cross-cultural knowledge communication and knowledge organization, philosophy of science, and information science.

### DANIEL SCHOPPER

Daniel Schopperis a junior scientist at the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences and head of its working group *Data, Resources and Standards*. Coming from a background of German philology and literature studies, he became involved with the digital humanities over the course of various edition projects, ranging from seventeenth-century drama to database-driven analysis of autobiographic writing, and is currently expanding his field of activity into linguistic-related areas. His research interests include interdisciplinary applications and methodologies between language corpora and literature studies, user–text interfaces, and web-based research environments.